Sur l'estimation du support d'une densité

Gérard Biau, Benoît Cadre, Bruno Pelletier

Sur l'estimation du support d'une densité

Gérard Biau [a] & Benoît Cadre [b] & Bruno Pelletier [c]

[a] *LSTA & LPMA*
Université Pierre et Marie Curie – Paris VI
Boîte 158, 175 rue du Chevaleret
75013 Paris, France
gerard.biau@upmc.fr

[b] *IRMAR, ENS Cachan Bretagne, CNRS, UEB*
Campus de Ker Lann
Avenue Robert Schuman
35170 Bruz, France
Benoit.Cadre@bretagne.ens-cachan.fr

[c] *IRMAR, Université Rennes 2, CNRS, UEB*
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France
bruno.pelletier@uhb.fr

Résumé

Etant donnée une densité de probabilité multivariée inconnue f à support compact et un n-échantillon i.i.d. issu de f, nous étudions l'estimateur du support de f défini par l'union des boules de rayon r_n centrées sur les observations. Afin de mesurer la qualité de l'estimation, nous utilisons un critère général fondé sur le volume de la différence symétrique. Sous quelques hypothèses peu restrictives, et en utilisant des outils de la géométrie riemannienne, nous établissons les vitesses de convergence exactes de l'estimateur du support tout en examinant les conséquences statistiques de ces résultats.

Mots-clés **— Estimation du support, géométrie riemannienne, statistique non paramétrique, théorème central limite, vitesses de convergence exactes, voisinage tubulaire.**

Abstract

Given an unknown multivariate probability density f with compact support, and an i.i.d. random n-sample drawn from f, we study the estimator of the support of f defined as unions of balls centered at the observations and of common radius r_n. To measure the quality of the estimation, we use a general criterion based on the volume of the symetric difference. Under some mild assumptions, and using tools from

Riemannian geometry, we establish the exact convergence rates of the estimator, and we discuss consequences of our results from a statistical perspective.

Index Terms **— Central limit theorem, nonparametric statistics, exact rates of convergence, Riemannian geometry, support estimation, tubular neighborhood.**

1 Introduction

Soit f une densité de probabilité inconnue définie par rapport à la mesure de Lebesgue λ sur $\mathbb{R}^d$. Etant donné un échantillon aléatoire $X_1, \ldots, X_n$ issu de f, nous nous intéressons dans ce travail à l'estimation du support de f, c'est-à-dire à l'ensemble fermé supposé compact

$$S_f = \overline{\{x \in \mathbb{R}^d : f(x) > 0\}}$$

(la barre supérieure désigne l'adhérence). Il s'agit d'un problème riche, qui trouve des applications dans des domaines aussi divers que le marketing, l'économétrie, le diagnostic médical ou encore le contrôle de qualité. Nous renvoyons par exemple le lecteur à Baíllo, Cuevas et Justel (2000) pour une introduction au sujet et une liste des références essentielles.

Parmi les nombreuses stratégies permettant d'estimer l'ensemble S_f, la plus simple est sans aucun doute celle choisie par Devroye et Wise (1980). L'estimateur proposé par ces auteurs est défini par

$$\hat{S}_n = \bigcup_{i=1}^{n} \mathcal{B}(X_i, r_n), \tag{1}$$

où $\mathcal{B}(x, r)$ désigne la boule euclidienne fermée centrée en x et de rayon r, et où (r_n) est une suite de nombres réels strictement positifs. Cette dernière suite joue le rôle d'un paramètre de lissage, analogue à celui de la fenêtre pour les estimateurs à noyau.

Afin d'apprécier la proximité entre l'estimateur $\hat{S}_n$ et la cible S_f, le critère le plus naturel est donné par la distance $d_1(\hat{S}_n, S_f)$, définie par

$$d_1(\hat{S}_n, S_f) = \lambda(\hat{S}_n \triangle S_f).$$

Ce critère peut être facilement étendu à d'autres distances, de la forme

$$d_\mu(\hat{S}_n, S_f) = \mu(\hat{S}_n \triangle S_f),$$

où μ désigne une mesure borélienne quelconque sur $\mathbb{R}^d$, éventuellement différente de la mesure de Lebesgue. Ainsi, et en supposant par exemple que μ admet une densité g par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, le critère d_μ se réécrit

$$d_g(\hat{S}_n, S_f) = \int_{\mathbb{R}^d} \mathbf{1}_{\hat{S}_n \triangle S_f}(x) g(x) \mathrm{d}x. \tag{2}$$

Cette mesure de proximité est très générale et permet, via le choix de g, d'adapter le critère d'erreur à l'objectif statistique. Par exemple, le choix $g \equiv f$ conduit à

$$d_f(\hat{S}_n, S_f) = \mathbb{P}(X \notin \hat{S}_n | X_1, \ldots, X_n),$$

où X désigne une variable aléatoire à densité f, indépendante de l'échantillon. Plus généralement, si X admet une densité g, nous obtenons

$$d_g(\hat{S}_n, S_f) = \mathbb{P}(X \in \hat{S}_n \triangle S_f | X_1, \ldots, X_n).$$

Cette dernière perte a été considérée par Devroye et Wise (1980) dans un problème de test visant à déterminer si une machine fonctionne dans des conditions normales ou pas. Ces auteurs ont en outre montré la convergence de l'estimateur $\hat{S}_n$ à l'aide du critère ci-dessus, moyennant des conditions sur la suite (r_n) analogues à celles que l'on impose traditionnellement en estimation de la densité.

2 Vitesses de convergence

A notre connaissance, aucune vitesse de convergence exacte relative à l'estimateur du support (1) n'est disponible dans la littérature. Dans ce travail, nous proposons donc d'analyser les vitesses de convergence de cet estimateur, en utilisant la distance d_g définie en (2) comme critère de qualité. Notre principal résultat exprime le fait que, sous certaines conditions analytiques peu contraignantes sur f et g, il existe une constante positive explicite c telle que

$$\sqrt{n r_n^d}\, \mathbb{E} d_g(\hat{S}_n, S_f) \to c \quad \text{lorsque } n \to +\infty,$$

à condition que $n r_n^d \to \infty$ et $n r_n^{d+2} \to 0$. En fait, nous montrons que bien d'autres types de vitesses sont possibles, qui dépendent essentiellement des positions relatives et des caractéristiques géométriques des supports de f et g.

L'originalité de notre travail réside en grande partie dans les outils de preuve que nous utilisons, qui sont pour l'essentiel issus de la géométrie riemannienne et de la théorie des voisinages tubulaires (Bredon, 1993). Nous terminons notre analyse en montrant que, sous certaines conditions peu contraignantes,

$$\left(\frac{n}{r_n^d}\right)^{1/4} \left(\lambda\left(\hat{S}_n \triangle S_f\right) - \mathbb{E}\lambda\left(\hat{S}_n \triangle S_f\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_f^2),$$

pour une certaine variance explicite σ_f^2. Ce théorème central limite, qui a été obtenu en collaboration avec D. Mason, répond en particulier à une conjecture laissée ouverte par P. Hall en 1985.

Les résultats présentés dans cette conférence correspondent aux publications [2] et [3].

Bibliographie

[1] Baíllo, A., Cuevas, A. et Justel, A. (2000). Set estimation and nonparametric detection, *Canadian Journal of Statistics*, Vol. 28, pp. 765–782.

[2] Biau, G., Cadre, B., Mason, D.M. et Pelletier, B. (2009). Asymptotic normality in density support estimation, *Electronic Journal of Probability*, Vol. 14, pp. 2617–2635.

[3] Biau, G., Cadre, B. et Pelletier, B. (2008). Exact rates in density support estimation, *Journal of Multivariate Analysis*, Vol. 99, pp. 2185–2207.

[4] Bredon, G.E. (1993). *Topology and Geometry*, Volume 139 of *Graduate Texts in Mathematics*, Springer-Verlag, New York.

[5] Devroye, L. et Wise, G. (1980). Detection of abnormal behavior via nonparametric estimation of the support, *SIAM Journal on Applied Mathematics*, Vol. 38, pp. 480–488.

[6] Hall, P. (1985). Three limit theorems for vacancy in multivariate coverage problems, *Journal of Multivariate Analysis*, **Vol. 16**, pp. 211-236.